METRUM
RESEARCH GROUP

Introduction to Bayesian pharmacometric data analysis using NONMEM®

Tim Waterhouse
Metrum Research Group

30 March 2023

Introduction to Bayesian pharmacometric data analysis using NONMEM®

- Why Bayesian?
- Introduction to Bayesian statistical principles and methods
  - Bayes Rule
  - Bayesian modeling & inference process
- Computation for Bayesian modeling
  - Maximum a Posteriori (MAP) Bayes
    - Individual: NONMEM POSTHOC
    - Population: Penalized Maximum Likelihood
  - Full Bayesian analysis
    - General computational approach: posterior simulation
    - Brief intro to Markov chain Monte Carlo (MCMC) simulation
      - Gibbs sampling
      - Metropolis-Hastings
      - Hamiltonian Monte Carlo and NUTS

Introduction to Bayesian pharmacometric data analysis using NONMEM®

- Overview of NONMEM® implementations
  - MAP estimation
    - Using prior distributions with optimization methods
  - MCMC: BAYES and NUTS methods
  - Prior specification in NONMEM®
- Prior distributions
  - Role of a prior distribution
  - Informative, uninformative or weakly informative?
- Model evaluation and comparison

Introduction to Bayesian pharmacometric data analysis using NONMEM®

- Assessing convergence and choosing numbers of burn-in and post-burn-in samples
- Getting your hands on posterior samples for individual parameters and predictions
- When stuff goes wrong
  - Diagnosing and remedying sampling problems encountered with MCMC
  - Reparameterization, e.g., centered vs non-centered parameterizations for hierarchical models
  - Prior distributions as part of the solution

Introduction to Bayesian pharmacometric data analysis using NONMEM®

- Practical strategies for selecting Bayesian estimation methods for specific types of problems
  - When to go Bayes (and why)?
  - Which method?
  - Which tool?
- What didn't we cover?

Why Bayesian?

- Pharmacometricians are often called on to leverage prior knowledge in order to interpret new data and facilitate decision-making in drug development.
  - Qualitative prior knowledge is captured in the mathematical form of a model, i.e., the **likelihood function**.
  - Quantitative prior knowledge may be captured in the form of probability distributions of model parameter values, i.e., **prior distributions**.
- Add **data** and you have all the ingredients of Bayesian data analysis.
- With Bayes Rule and suitable computation tools those components are combined to yield **posterior distributions** of model parameters and predictions.
- Those distributions permit probabilistic inferences directly relevant to decision-making.

Why Bayesian analysis for pharmacometrics applications?

- Decision-making supported by quantitative synthesis of prior knowledge and heterogenous data.
- Calibration (and recalibration) of complex QSP models as new data accumulates.
- Bayesian framework more easily accommodates
  - model complexity, particularly in the stochastic structure of a model,
  - analysis of data from heterogeneous sources.

Introduction to Bayesian statistical principles and methods

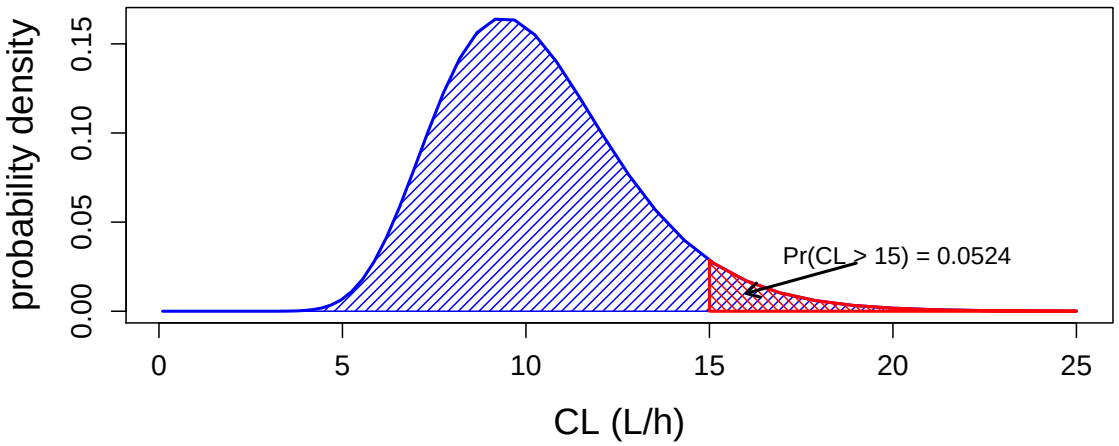
- Bayesian principles and methods provide a coherent framework for:
  - Quantifying uncertainty,
  - Making inferences in the presence of that uncertainty.
- It is also the basis for formal approaches to incremental model building, parameter estimation and other statistical inference as knowledge and data are accumulated.

The two core notions that distinguish Bayesian analysis are:

- Unknown quantities are viewed as random variables, i.e., they are described in terms of probability distributions.
- Bayes rule provides a formal mechanism for combining prior knowledge and new data.

Bayesian approach to uncertainty

- Uncertainty is described in terms of probability:
  - Can assign a quantitative description of uncertainty in the model parameters and evaluate how it affects the uncertainty in predicted outcomes.
  - Probability can be used to quantitatively represent uncertainty based on either “objective” evidence/data or “subjective” expert opinion or belief.



Bayesian approach to statistical inference

- Model parameters and predictions are described in terms of probability distributions representing uncertainty.
- Results reflect the combined evidence of data and prior knowledge or belief.
- Focuses on estimation and inferences related to probabilities of unknown quantities: parameters, future data, hypotheses.
- Inferences may be described directly in terms of probabilities, e.g.,
  - What is the probability that a parameter or a function of parameters is within or outside of a specific interval?
  - What is the lowest dose that will achieve a $\geq$ 80% probability of reaching an efficacy or safety target?

Bayesian inference
Bayes Rule

Bayes Rule is the basis for inference about model parameters (θ) given data (y) and prior knowledge about model parameters ($p(\theta)$):

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)\,p(y|\theta)}{p(y)} = \frac{p(\theta)\,p(y|\theta)}{\int p(\theta)\,p(y|\theta)\,d\theta} \\ &\propto p(\theta)\,p(y|\theta) \end{aligned}$$

The p's are probabilities or probability densities of the specified random variables.

Bayesian modeling/inference process

1. Assess prior distribution $p(\theta)$
   - θ viewed as random variables
   - Subjective
   - Ideally base on all available evidence/knowledge (or belief)
   - Or deliberately select a non-informative prior (e.g., reference, vague or improper prior)
2. Construct a model for the data $p(y|\theta)$, also known as the likelihood function when viewed as a function of θ.
3. Calculate posterior distribution $p(\theta|y)$.
   - Use for inferences regarding parameter values
4. Calculate posterior predictive distribution $p(y_{\text{new}}|y)$.
   - Use for inferences regarding future observations

$$p(y_{new}|y) = \int p(y_{new}|\theta)\,p(\theta|y)\,d\theta$$

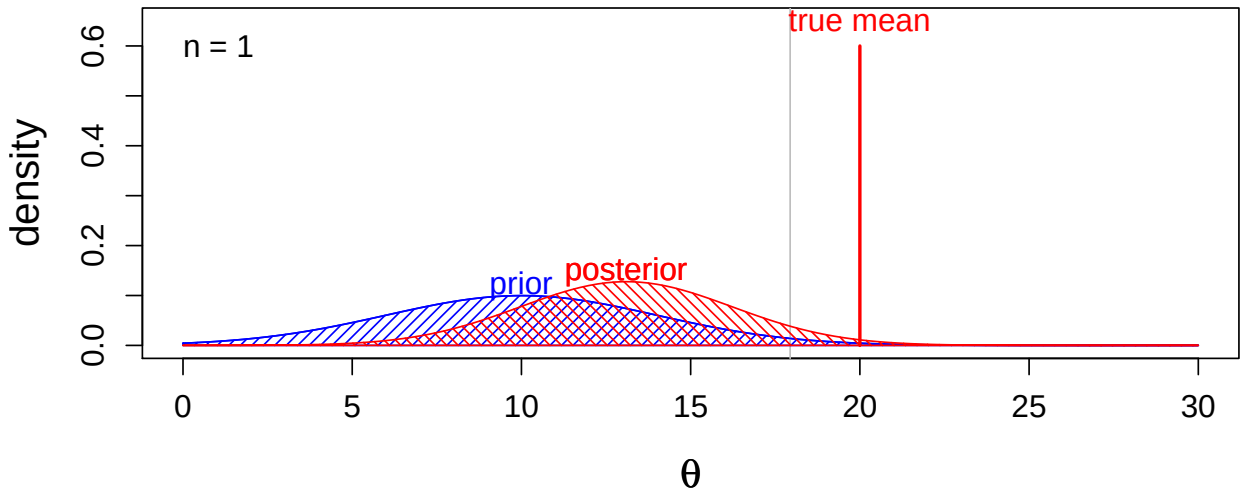
A simple one parameter example

Estimating the mean of a normal distribution with a known variance where the prior is also normal:

$$
\begin{aligned}
\theta &\sim N(\mu_0, \tau_0) \qquad\qquad y|\theta \sim N(\theta, \sigma) \\
p(\theta|y) &\propto p(\theta)\, p(y|\theta) = p(\theta) \prod_{i=1}^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\right) \\
&\Downarrow \\
\theta|y &\sim N(\mu_n, \tau_n) \\
\mu_n &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}
\end{aligned}
$$

A simple one parameter example

- Posterior mean is a weighted average of prior mean (μ_0) and the sample mean ($\overline{y}$).
- Posterior precision $\left(\frac{1}{\tau_n^2}\right)$ is a sum of the prior precision $\left(\frac{1}{\tau_0^2}\right)$ and the data precision $\left(\frac{n}{\sigma^2}\right)$.
- The prior can also be interpreted in terms of equivalent number of data points, i.e., $n_0 = \frac{\sigma^2}{\tau_0^2}$.



Computation for Bayesian modeling

- Full Bayesian analysis requires:
  - Characterization of the joint posterior distribution of model parameters and of predicted outcomes.
  - Integration of that joint posterior distribution to calculate quantities required for statistical inferences.
- For most realistic problems, those are very computationally demanding tasks.
- Increases in computation speed and development of new algorithms over the last 25–30 years have finally made full Bayesian analysis a feasible option for routine data analysis.

Maximum a Posteriori (MAP) Bayes

- The set of model parameter values that maximize the posterior distribution (aka posterior mode) are called the MAP Bayes parameter estimates.
- MAP Bayes estimation is a useful approximate Bayesian method.
- It is strongly analogous to maximum likelihood (ML) estimation.
- ML estimation tools can often be adapted to calculate MAP Bayes estimates.

Full Bayesian analysis

- Full Bayesian analysis refers to characterization of the joint posterior distribution, not just its mode.
- Most Bayesian inference is based on univariate marginal posterior distributions of individual parameters or functions of those parameters.
- That requires integration of the joint posterior distribution, usually over several dimensions.

Full Bayesian analysis

- If you're lucky those integrals have known analytic solutions, but that is rarely true for PK/PD modeling applications.
- For integrals in fewer dimensions, a numerical quadrature method might be practical.
- Now imagine the computational requirements for hierarchical models, e.g., population PK models, with individual-specific parameters in the hundreds!!

Posterior simulation

- What if you could simulate samples of θ from the joint posterior distribution?
- Then you could estimate $E\left(f\left(\theta\right)|y\right)$ by the arithmetic mean:

$$E\left(f\left(\theta\right)|y\right) \approx \frac{1}{n}\sum_{i=1}^{n} f\left(\theta_i\right)$$

- More generally, you could characterize the properties of any marginal posterior distribution of a model parameter or function of model parameters, e.g., moments, quantiles, ...
- But how do you simulate samples from a high dimensional joint posterior distribution?
- Markov chain Monte Carlo (MCMC) simulation via NONMEM is one approach we will explore today.

Inferences from posterior simulations

- Posterior simulation yields vectors of parameters and/or predictions from a joint posterior distribution
- Marginal distributions
  - To describe the posterior distribution of any scalar function of the parameters apply the function to each simulated vector. The empirical distribution of those values approximates the posterior distribution.
  - The marginal distribution of any single parameter is just a special case of that approach.
- Inferences are usually based on moments, probabilities or percentiles from marginal posterior distributions. They are readily estimated from the corresponding sample statistics for the simulated values.

Markov Chain Monte Carlo (MCMC) simulation

- Involves random draws from approximate distributions and then correcting those draws to better approximate the joint posterior.
- The samples are drawn sequentially so that each draw depends on the previous one, thus forming a Markov chain.
- Eventually the Markov chain converges (in distribution) to a stationary distribution that is the joint posterior distribution.
- Algorithms for MCMC include:
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Hamiltonian Monte Carlo (HMC) simulation
- MCMC samples are serially correlated:
  - Inferences based on MCMC require more samples than would be required for independent samples
- Practical consequences:
  - Use only samples drawn after convergence is achieved, i.e., discard samples from a “warmup” phase.
  - Draw more samples than you would for independent random draws.

Gibbs sampling

- In most cases a convergent Markov chain may be constructed by progressively sampling from the univariate full conditional distributions:

$$
\begin{aligned}
\theta_1^i &\sim p\left(\theta_1 | \theta_2^{i-1}, \theta_3^{i-1}, \ldots, \theta_n^{i-1}, y\right) \\
\theta_2^i &\sim p\left(\theta_2 | \theta_1^{i}, \theta_3^{i-1}, \ldots, \theta_n^{i-1}, y\right) \\
&\vdots \\
\theta_n^i &\sim p\left(\theta_n | \theta_1^{i}, \theta_2^{i}, \ldots, \theta_{n-1}^{i}, y\right)
\end{aligned}
$$

- This reduces the multivariate posterior sampling problem to a sequence of more manageable univariate sampling problems.

Metropolis-Hasting algorithm

- The Metropolis-Hastings is a general purpose multivariate MCMC algorithm.
- It requires selection of a conditional proposal density $q(y|x)$ that is easy to sample from.
- To generate $x^{(t+1)} \sim f$ given a previous value $x^{(t)}$:

1. Generate $y_t \sim q\left(y|x^{(t)}\right)$.
2. $x^{(t+1)} = \begin{cases} y_t, & \text{with probability } \rho\left(x^{(t)}, y_t\right) \\ x^{(t)}, & \text{with probability } 1 - \rho\left(x^{(t)}, y_t\right) \end{cases}$
   where $\rho(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$

Hamiltonian Monte Carlo (HMC) simulation

Physical analogy to motivate HMC

- In classical mechanics the Hamiltonian equations describe the evolution of a system over time.
- The state of the system is described in terms of kinetic energy as a function of momentum (mass $\times$ velocity) and potential energy as a function of position.
- For the analogy equate the model parameters θ to position and equate a set of auxiliary parameters ρ to momentum.
- Now define a Hamiltonian in terms of the joint posterior distribution of θ and ρ:

$$\begin{aligned} H(\theta, \rho) &= -\log\left(p\left(\theta, \rho|y\right)\right) = -\log\left(p\left(\theta|y\right)p\left(\rho|\theta, y\right)\right) \\ &= -\log\left(p\left(\theta|y\right)\right) - \log\left(p\left(\rho|\theta, y\right)\right) \\ &= V\left(\theta\right) + T\left(\rho|\theta\right) \\ V\left(\theta\right) &= -log\left(p\left(\theta|y\right)\right) = \text{potential energy} \\ T\left(\rho|\theta\right) &= -log\left(p\left(\rho|\theta, y\right)\right) = \text{kinetic energy} \end{aligned}$$

Hamiltonian Monte Carlo (HMC) simulation

- θ is what we really care about.
- ρ allows the use of Hamiltonian mechanics to more efficiently move through the relevant parts of the parameter space.
- Usually the distribution of ρ is chosen to be independent of θ, e.g., $p(\rho|\theta) = p(\rho) = N(0, \Sigma)$.
- Suppose we place a frictionless particle on the potential energy surface $(-log(p(\theta|y)))$ at some position θ^{t-1}.
  - We give it a shove that imparts a momentum ρ^{t-1} to that particle at time $t-1$.
  - The particle moves over that surface according to Hamiltonian dynamics.
  - Now stop the particle at time t and measure its position θ^t.
  - Now randomly sample a new momentum from $p(\rho)$ and give the particle another shove, and so on...

Hamiltonian Monte Carlo (HMC) simulation

- Though the initial momentum at each step is random, the subsequent path will favor regions of lower potential energy (higher probability density).
- The set of sampled positions are distributed according to the target posterior density.
- In practice the Hamiltonian equations are solved numerically. As a result some error is introduced in the estimated path.
- A Metropolis step is used to assure that the position samples converge in distribution to the target distribution.

The HMC algorithm

Repeat the following steps:

1. Sample $\rho^{t-1} \sim N(0, \Sigma)$
2. Simultaneously update θ and ρ by numerically solving the Hamiltonian equations using the leapfrog method to generate a proposal θ^* for θ^t.
3. Apply a Metropolis step to decide whether to accept or reject the proposal θ^* as θ^t.

The leapfrog method

Using the starting values θ^{t-1} and ρ^{t-1} the leapfrog algorithm alternates half-step updates of ρ with full step updates of θ:

$$\rho \leftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta} = \rho + \frac{\epsilon}{2}\frac{d \log\left(p\left(\theta|Y\right)\right)}{d\theta}$$
$$\theta \leftarrow \theta + \epsilon \Sigma \rho$$
$$\rho \leftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta} = \rho + \frac{\epsilon}{2}\frac{d \log\left(p\left(\theta|Y\right)\right)}{d\theta}$$

For each HMC iteration repeat this L times to yield the proposal values θ^* and ρ^*.

The Metropolis step

- Compute the ratio:

$$r = \exp\left(H\left(\theta^{t-1}, \rho^{t-1}\right) - H\left(\theta^{*}, \rho^{*}\right)\right)$$
$$= \frac{p\left(\theta^{*}|y\right) p\left(\rho^{*}\right)}{p\left(\theta^{t-1}|y\right) p\left(\rho^{t-1}\right)}$$

- Accept/reject step:

$$\theta^{t} = \begin{cases} \theta^{*}, & \text{with probability } \min\left(r, 1\right) \\ \theta^{t-1}, & \text{otherwise} \end{cases}$$

- Since ρ is sampled independently of θ and previous values of ρ, we just discard ρ^{*} and sample a new value for the next HMC iteration.

HMC algorithm parameters

Parameters that must be set: discretization time ϵ, number of leapfrog steps L and mass matrix Σ^{-1}.

Sampling efficiency is very sensitive to those parameters:

- ϵ too large $\rightarrow$ too many proposals rejected
- ϵ too small $\rightarrow$ long simulation times
- L too large $\rightarrow$ too much work for each iteration
- L too small $\rightarrow$ devolves to a random walk
- If Σ^{-1} is poorly tuned to the problem, ϵ needs to be decreased and L increased to maintain precision and efficiency.

Stan automatically optimizes those parameters using the NUTS (no U-turn sampling) algorithm [1].

HMC performance

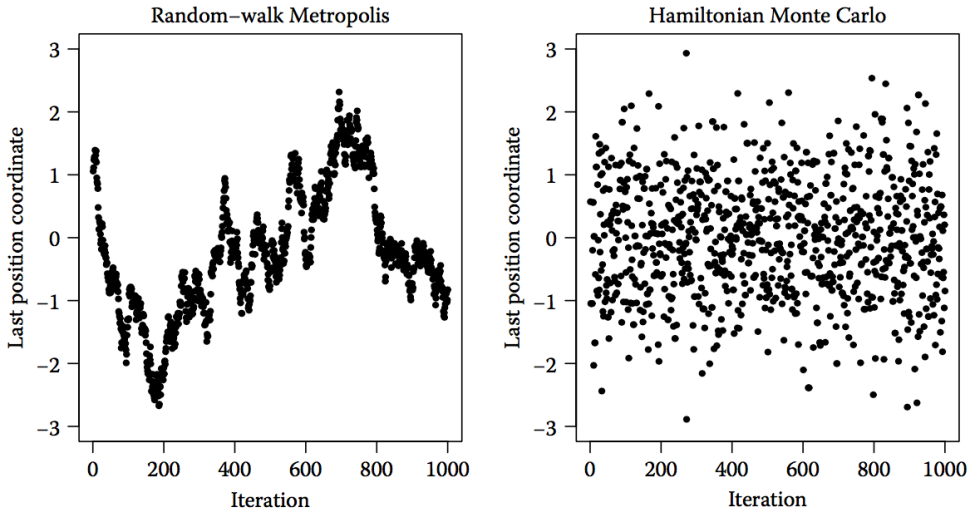


FIGURE 5.6
Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.

from RM Neal. MCMC Using Hamiltonian Dynamics (2011) [2]

HMC performance

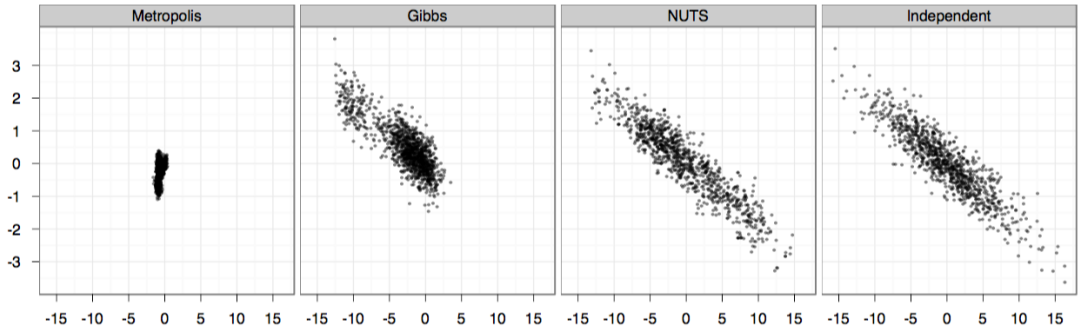


Figure 7: *Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000 samples (thinned to 1,000 samples for display) generated by Gibbs sampling (second from left), and 1,000 samples generated by NUTS (second from right). Only the first two dimensions are shown here.*

from MD Hoffman and A Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo (2014) [1]

HMC issues/limitations

- Requires calculation of the gradient $\frac{d \log(p(\theta|Y))}{d\theta}$
- Suitable for sampling of continuous parameters only
  - Cannot sample discrete parameters
  - Discrete data is OK as long as the likelihood depends only on continuous parameters.
  - Models with discrete parameters, e.g., finite mixture models, can often be implemented by marginalizing out the discrete parameters.

Overview of NONMEM implementations

- MAP estimation
  - Using prior distributions with any optimization method
- MCMC
  - METHOD = BAYES: Metropolis-Hastings within Gibbs sampling
  - METHOD = NUTS: No U-turn sampler (HMC with automatic optimization of sampling parameters)

Prior specification in NONMEM

- THETA: (multivariate) normal distribution
- OMEGA & SIGMA: inverse Wishart
  - NUTS also supports
    - lognormal or half-t distributions for SDs, i.e., square roots of the diagonal elements of OMEGA or SIGMA
    - LKJ distribution for the correlation matrix

Brief discussion on prior distributions

- Think of prior distributions as part of the model.
- Priors should be chosen and subjected to scrutiny much like other model components.
- Model checking should ideally include sensitivity analysis of the priors.
- Choice of priors is most critical with sparse or limited data.

See https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

What is the function of a prior distribution

- Represent prior knowledge
- Regularization to facilitate computation
  - Typically weakly-moderately informative
  - E.g., Cauchy with most of its mass in a plausible range, but heavy tails allow for diagnosis of prior-posterior discrepancies.

What does it mean to be informative, uninformative or weakly informative?

Not well defined, but here's an attempt at some loose definitions:

- Weakly informative prior: A prior that rules out unreasonable parameter values but is not so strong as to rule out values that might make sense
- Informative prior: A prior that purposely represents information intended to influence the posterior distribution
  - To capture prior knowledge
  - To challenge the analysis with competing points of view, e.g., use of pessimistic or optimistic priors.
- Uninformative prior: Ostensibly a prior that represents no information and therefore "let's the data tell the story."
  - E.g., a constant over the entire real line—an improper prior

Beware: That “uninformative” prior might not be!

- Suppose you use an improper prior for a standard deviation—a constant over the positive real line.
- That means all positive values are equally likely. Sounds like a reasonable definition of uninformative doesn’t it?
- But that means that the prior assigns infinitely more probability to the set of values greater than any fixed value you care to choose.
- This will tend to bias the posterior to high values.
- Bottom line: A uniform distribution does not automatically confer uninformativeness.

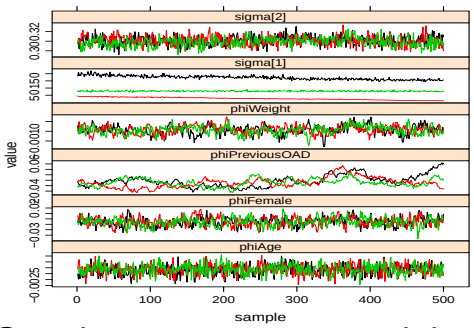
Assessing convergence & adequacy of sample sizes

- Early samples may be unrepresentative of the target distribution
- MCMC samples within a chain are autocorrelated
  - Inferences based on MCMC samples are less precise than those from the same number of independent samples
  - Autocorrelation also influences the rate of convergence

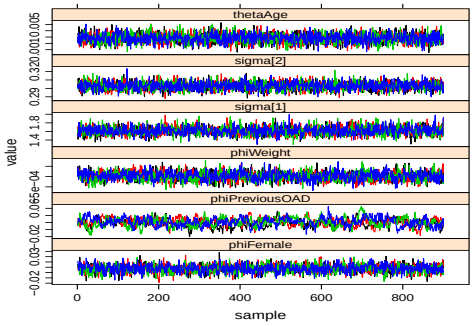
Assessing convergence & adequacy of sample sizes

- Use a warmup phase, i.e., discard early iterations
- Monitor convergence via multiple chains with different starting points
  - Look for chains to converge to a common distribution
  - You want chain history plots to look more like straight horizontal “fuzzy caterpillars” than “wiggly snakes”
  - Monitor Gelman-Rubin diagnostics (Rhat) and/or Gelman-Rubin-Brooks plots
    - Essentially ratios of total variance to within chain variance.
    - Should approach 1 for all parameters of interest on convergence

Poor convergence & mixing



Good convergence & mixing



IMHO

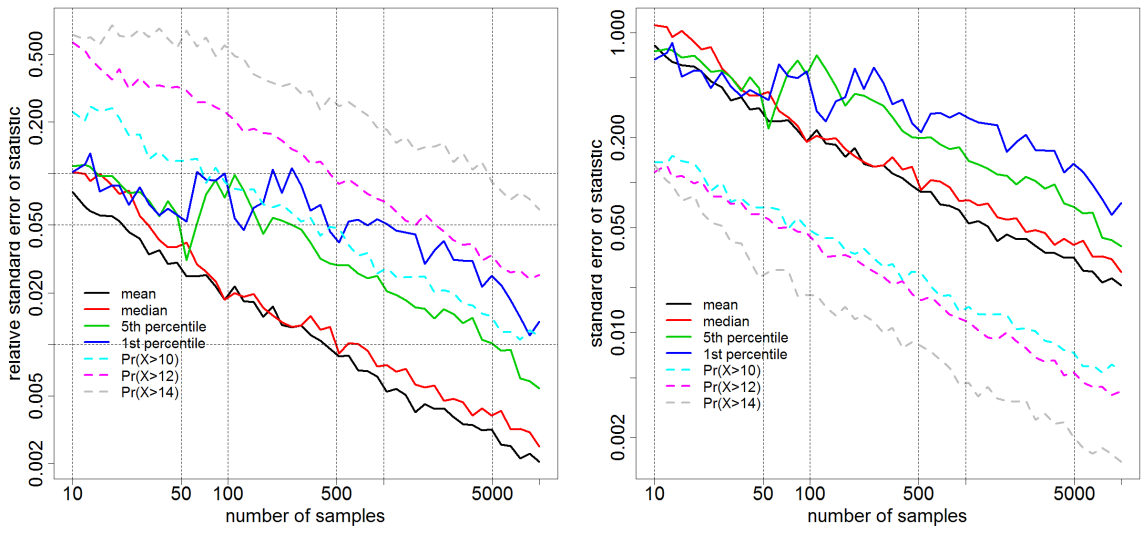
- Do NOT trust automatic convergence detection to determine number of burn-in samples.
- In particular, do NOT use NONMEM’s termination tests (CTYPE option) for terminating burn in.

How many samples?

- Number of samples depends on the inference(s) of interest and the desired precision
- For independent samples:
  - A posterior mean of a parameter θ is estimated with an error of $\sim \frac{s_\theta}{\sqrt{n}}$ where s_θ is the standard deviation of the simulated values of θ and n is the number of samples.
  - A probability p is estimated with an error of $\sim \sqrt{\frac{p(1-p)}{n}}$.
- In general more samples are required for estimating tail quantiles and probabilities than for central tendencies and probabilities near 0.5

How many samples?

- Suppose the posterior distribution of θ is $N(\mu = 10, \sigma = 2)$.
- What if we use simulation to estimate various features of that distribution?
- The following are bootstrap estimates of error due to simulation:



How many MCMC samples?

- Given a set of MCMC samples it is possible to adjust for autocorrelation to estimate:
  - The equivalent number of independent samples, aka effective sample size (n_eff)
  - The standard error in the estimated posterior mean
  - The rstan print and summary functions provide estimates of each.
- Guidance based on independent samples may then be applied

For more rigorous and comprehensive treatment see Robert and Casella 2004 and 2010 [5, 6].

Getting your hands on posterior samples for individual parameters and predictions

- With METHOD = BAYES or NUTS NONMEM writes MCMC sampled parameter values to a file named in the $ESTIMATION statement (Defaults to <model name>.ext).
- Only population parameters are written.
- Before NONMEM 7.5: Use FORTRAN WRITE statements (verbatim code) to write individual parameter or prediction values to files.
- NONMEM 7.5: Use $EST ... BAYES_PHI_STORE=1 to generate <model name>.iph

Priors for covariance matrices (OMEGA & SIGMA)

- METHOD = BAYES
  - Inverse Wishart
- METHOD = NUTS
  - OLKJDF = 0: Inverse Wishart
  - OLKJDF $>$ 0: LKJ distribution for the correlation matrix with shape parameter = OLKJDF
    - OVARF $>$ 0: lognormal distribution for $\sqrt{\Omega_{ii}}$ where Ω_{ii} is a diagonal element of OMEGA parameterized as:

$$\log(\sqrt{\Omega_{ii}}) \sim N\left(\log\left(\sqrt{\text{OmegaPrior(i,i)}}\right), \frac{1}{\text{OVARF}}\right).$$

  - OVARF $<$ 0: half-t distribution for $\sqrt{\Omega_{ii}}$ where Ω_{ii} is a diagonal element of OMEGA parameterized as:

$$\sqrt{\Omega_{ii}} \sim \text{half-}t\left(0, \sqrt{\text{OmegaPrior(i,i)}}, |OVARF|\right)$$

NONMEM parameterization of the inverse Wishart distribution

NONMEM implementation of inverse Wishart prior for Ω (or Σ):

$$\begin{aligned}
\Omega &\sim W^{-1}\left(\nu\Omega_{\text{prior}}, \nu\right) \\
W^{-1} &= \text{Standard parameterization of the inverse Wishart} \\
\nu &= \text{degrees of freedom} \\
E(\Omega) &= \frac{\nu\Omega_{\text{prior}}}{\nu - n - 1} \\
\text{Var}(\Omega_{ii}) &= \frac{E(\Omega_{ii})(\nu - n - 1)}{\nu} \\
n &= \dim(\Omega)
\end{aligned}$$

Inverse Wishart distribution

Guidance for setting $Omega_{\text{prior}}$ and ν:

$$\nu_i = \frac{2E\left(\Omega_{ii}\right)^2}{\text{Var}\left(\Omega_{ii}\right)} + n + 3$$
$$\Omega_{\text{prior},ii} = \frac{E\left(\Omega_{ii}\right)\left(\nu_i - n - 1\right)}{\nu_i}$$

- Set diagonal elements of Ω_{prior} to the calculated values of $\Omega_{\text{prior},ii}$
- Set $\nu = \min\left(\nu_i\right)$.

When stuff goes wrong

- Diagnosing and remedying sampling problems encountered with MCMC
- Reparameterization, e.g., centered vs non-centered parameterizations for hierarchical models
- Prior distributions as part of the solution

Improving computational efficiency and sampling performance

The main strategies are:

- Reparameterization
- Adjusting MCMC tuning parameters
- Weakly informative priors to regularize fitting of hierarchical models

Reparameterization

- Model parameterization can markedly affect the geometry of the joint posterior distribution.
- Posterior distributions with extreme curvature or sharp transitions lead to sampling inefficiency or outright failure.
- A well-selected reparameterization can often “smooth out” the geometry, thereby improving sampling performance.
- For example switch from centered to non-centered parameterization (or vice versa) of random effects in hierarchical models.
- Other reparameterizations that reduce posterior curvature and correlation, e.g., truncated Emax parameterization.
- Use MU referencing when possible with BAYES and NUTS.

Centered (CP) and non-centered (NCP) parameterization with NONMEM

Consider an individual parameter $\phi_i \sim N(\mu, \omega)$.

- CP refers to specifying ϕ_i directly using that distribution centered at μ.
- NCP refers to specifying ϕ_i indirectly using a standard normal, $\eta_{\text{std},i} \sim N(0, 1)$, and then calculating $\phi_i = \mu + \omega\eta_{\text{std},i}$
- The above is generalized to multivariate normal case using a Cholesky decomposition to generate a multivariate normal vector from standard normal ηs.

Centered (CP) and non-centered (NCP) parameterization

- For NUTS,
  - AUTO = 0 or 1 for CP
  - AUTO = 2 for NCP
  - AUTO = 3 for something in between. Uses $N(0, \omega)$ instead of $N(0, 1)$
- To my knowledge there are no equivalent implementations for BAYES.

Reparameterize to reduce correlation among parameters

PII-108

"TRUNCATED SIGMOID E_{max} MODELS": A REPARAMETERIZATION OF THE SIGMOID E_{max} MODEL FOR USE WITH TRUNCATED PK/PD DATA. WJ Bachman PhD and WR Gillespie PhD, GloboMax LLC, Hanover, MD.

The parameters of the sigmoid E_{max} model are poorly estimated when the range of PK/PD data available is limited to $<0.95E_{max}$ [Dutta et al. J Pharm Sci 85:232 (1996)]. The following reparameterized form of the sigmoid E_{max} model has improved parameter estimation properties:

$$E = E_0 + \frac{(\beta^\gamma + 1)(E^* - E_0)C^\gamma}{C^{*\gamma} + \beta^\gamma C^\gamma}$$

where E is the effect measure and C is a measure of drug exposure (e.g., concentration or dose). The parameter E^* is the estimated effect resulting from C^*, γ is the usual "sigmoidicity" parameter, and E_0 is the baseline effect. β is a measure of the degree to which the function deviates from linearity in C^γ. One approach to applying this parameterization is to fix C^* (or E^*) at a value and estimate the remaining parameters E_0, E^* (or C^*), β, and γ by nonlinear regression. The properties of this approach are evaluated by application to simulated PK/PD data that is truncated at various fractions of E_{max}. When C^* (or E^*) is chosen within the range of the observed data, then the parameters E^* (or C^*) and β are more precisely and accurately estimated than EC_{50} and E_{max} of the standard parameterization.

CP&T 63:199 (1998)

- With BAYES method high correlation among parameters often causes high autocorrelation in the MCMC samples.
- This often happens with asymptotic functions like the Emax function.
- NUTS is more robust to such correlation.

Prior distributions as part of the solution

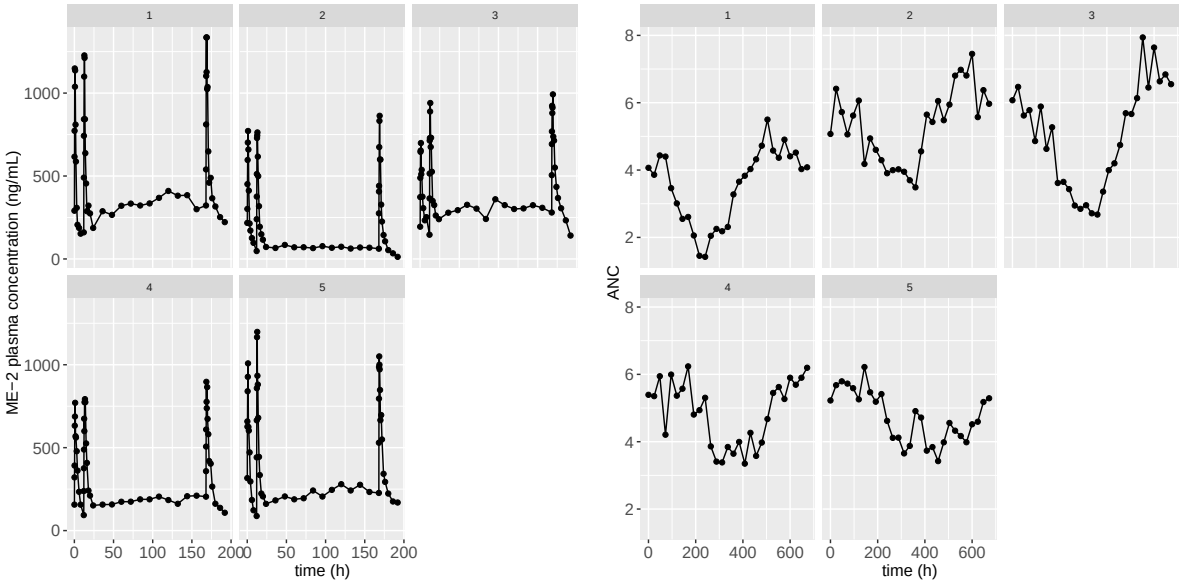
IIV variances are difficult to estimate, particularly with data from a small number of individuals. What to do?

- Reduce the number of random effects until you find a set you can estimate with high precision, or
- Use a weakly informative prior for Ω that is consistent with our knowledge of IIV and excludes clearly implausible values.

I argue that for most PMX models the latter is more consistent with our knowledge and should be the preferred approach.

Example 4: Full Bayes popPKPD using semi-mechanistic model

PKPD model of drug-induced neutropenia

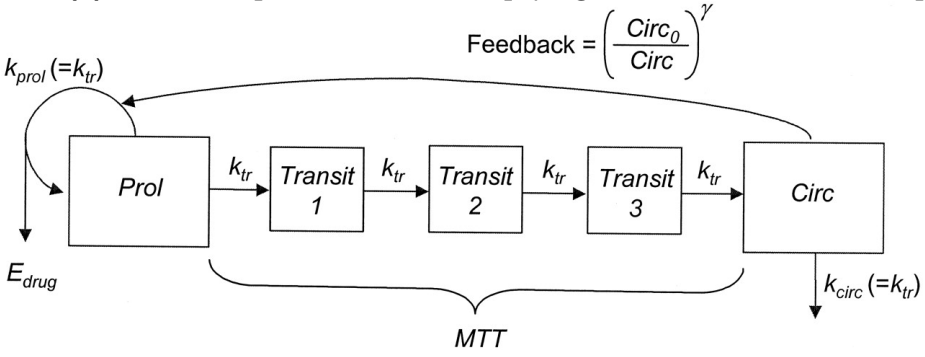


Friberg-Karlsson semi-mechanistic model for drug-induced myelosuppression

- PK model: Two compartment model with first order absorption describing plasma drug concentration on the i^{th} occasion in the j^{th} subject as a function of time, dose and body weight:

$$\begin{aligned} \log\left(c_{ij}\right) &\sim N\left(\log\left(\hat{c}_{ij}\right), \sigma\right) \\ \hat{c}_{ij} &= f_{2cpt}\left(t_{ij}, D_j, \tau_j, CL_j, Q_j, V_{1j}, V_{2j}, k_{aj}\right) \end{aligned}$$

- Friberg-Karlsson semi-mechanistic model for drug-induced myelosuppression [7, 8, 9, 10, 11] (Figure 2 of reference [7])



Friberg-Karlsson semi-mechanistic model for drug-induced myelosuppression

$$
\begin{aligned}
\frac{dProl}{dt} &= k_{prol} Prol\,(1 - E_{drug}) \left(\frac{Circ_0}{Circ}\right)^{\gamma} - k_{tr} Prol \\
\frac{dTransit1}{dt} &= k_{tr} Prol - k_{tr} Transit1 \\
\frac{dTransit2}{dt} &= k_{tr} Transit1 - k_{tr} Transit2 \\
\frac{dTransit3}{dt} &= k_{tr} Transit2 - k_{tr} Transit3 \\
\frac{dCirc}{dt} &= k_{tr} Transit3 - k_{circ} Circ
\end{aligned}
$$

$\widehat{c}$ $\equiv$ plasma drug concentration

$Circ$ $\equiv$ absolute neutrophil count (ANC)

Parameters in red are *system* parameters, i.e., drug-independent.

$$
\begin{aligned}
E_{drug} &= \alpha \widehat{c} \\
k_{prol} &= k_{circ} = k_{tr} \\
MTT &= \frac{n+1}{k_{tr}}
\end{aligned}
$$

Friberg-Karlsson model

$$
\begin{aligned}
\log(c_i) &\sim N(\log(\hat{c}_i), \sigma_{PK}) \\
\hat{c}_i &= \frac{x_{2i}}{V_1} \\
x_1' &= -k_a x_1 \\
x_3' &= \frac{CL+Q}{V_1} x_2 - \frac{Q}{V_2} x_3 \\
x_5' &= k_{tr}(\text{Prol} - \text{Transit}_1) \\
x_7' &= k_{tr}(\text{Transit}_2 - \text{Transit}_3)
\end{aligned}
$$

where

$$
\begin{aligned}
x_4 &= \text{Prol} \\
x_6 &= \text{Transit}_2 \\
x_8 &= \text{Circ} \\
x(0) &= \{0, 0, 0, \text{Circ}_0, \text{Circ}_0, \text{Circ}_0, \text{Circ}_0, \text{Circ}_0\}
\end{aligned}
$$

$$
\begin{aligned}
\log(ANC_i) &\sim N(\log(\text{Circ}_i), \sigma_{PD}) \\
\text{Circ}_i &= x_{8i} \\
x_2' &= k_a x_1 - \frac{CL+Q}{V_1} x_2 + \frac{Q}{V_2} x_3 \\
x_4' &= k_{tr} \text{Prol} \left((1 - E_{\text{drug}}) \left(\frac{\text{Circ}_0}{\text{Circ}}\right)^{\gamma} - 1\right) \\
x_6' &= k_{tr}(\text{Transit}_1 - \text{Transit}_2) \\
x_8' &= k_{tr}(\text{Transit}_3 - \text{Circ})
\end{aligned}
$$

$$
\begin{aligned}
x_5 &= \text{Transit}_1 \\
x_7 &= \text{Transit}_3 \\
k_{tr} &= \frac{4}{MTT}
\end{aligned}
$$

IIV and prior distributions

- Inter-individual variation

$$\log\left(CL_j, Q_j, V_{1j}, V_{2j}, k_{aj}, MTT_j, Circ_{0j}, \alpha_j\right) \sim N\left(\log\left(\widehat{CL}\left(\frac{bw_j}{70}\right)^{0.75}, \widehat{Q}\left(\frac{bw_j}{70}\right)^{0.75}, \widehat{V}_1\left(\frac{bw_j}{70}\right), \widehat{V}_2\left(\frac{bw_j}{70}\right), \widehat{k}_a, \widehat{MTT}, \widehat{Circ_0}, \widehat{\alpha}\right), \Omega\right)$$

- Prior distributions: moderately informative for PK, strongly informative for system parameters, weakly informative for drug effect

$$\widehat{CL} \sim \log N(\log(10), 0.5) \quad \widehat{Q} \sim \log N(\log(15), 0.5) \quad \widehat{V}_1 \sim \log N(\log(35), 0.5)$$
$$\widehat{V}_2 \sim \log N(\log(105), 0.5) \quad \widehat{k}_a \sim \log N(\log(2), 0.5)$$
$$\widehat{MTT} \sim \log N(\log(125), 0.2) \quad \widehat{Circ_0} \sim \log N(\log(5), 0.2) \quad \gamma \sim \log N(\log(0.17), 0.2)$$
$$\widehat{\alpha} \sim \log N\left(\log(3 \times 10^{-4}), 1\right)$$
$$\Omega \sim \text{inverse Wishart}\left(\begin{bmatrix} 0.045 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.045 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.045 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.045 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.045 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.045 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.045 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.045 \end{bmatrix}, 11\right)$$

Practical strategies for selecting Bayesian estimation methods for specific types of problems

- When to go Bayes (and why)?
- Which method?
- Which tool?

When to go Bayes (and why)?

- Bayesian data analysis and inference are always applicable to PMX modeling applications.
  - I think most scientists tend to interpret probability in Bayesian terms, so it makes sense to use methods consistent with that.
  - However, BDA is usually more computationally demanding and time consuming.
- Pragmatism leads us to go Bayesian when it pays off the most.
- E.g., when you want your inferences to be informed by both prior information and new data.
- IMHO we should be using informative prior distributions more extensively, particularly during the “learn” stages of drug development.

Which method?

- MCMC when it is practical.
- MAP Bayes for large scale models (e.g., QSP) that require more computation time than is practical for MCMC.
- MAP Bayes for more rapid exploration of multiple models followed by MCMC for the final model (or small set of contenders).
- Generally prefer NUTS over MH/Gibbs (BAYES)

Further reading

- Introduction to Bayesian pharmacometric data analysis with NONMEM (ACoP 2019 workshop by Bill Gillespie and Curtis Johnston)
- Bauer RJ. NONMEM Tutorial Part II: Estimation Methods and Advanced Examples. *CPT Pharmacometrics Syst Pharmacol*. 2019;8(8):538-556. doi:10.1002/psp4.12422
- Supplementary code for NONMEM Bayes tutorial paper (submitted): `https://github.com/metrumresearchgroup/NMBayesTutorial`

References I

[1] Matthew D Hoffman and Andrew Gelman.
The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.
The Journal of Machine Learning Research, 15(1):1593–1623, 2014.

[2] Radford M. Neal.
MCMC using hamiltonian dynamics.
In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman & Hall/CRC, Boca Raton, FL, 2011.

[3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin.
Bayesian Data Analysis.
CRC Press, Boca Raton, FL, third edition, 2014.

[4] Aki Vehtari, Andrew Gelman, and Jonah Gabry.
Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models.
arXiv preprint arXiv:1507.04544, 2015.

[5] Christian P. Robert and George Casella.
Monte Carlo Statistical Methods.
Springer, second edition, 2004.

References II

[6] Christian P. Robert and George Casella.
Introducing Monte Carlo Methods with R.
Springer, 2010.

[7] L E Friberg, A Henningsson, H Maas, L Nguyen, and M O Karlsson.
Model of chemotherapy-induced myelosuppression with parameter consistency across drugs.
J. Clin. Oncol., 20(24):4713–4721, 2002.

[8] L E Friberg and M O Karlsson.
Mechanistic models for myelosuppression.
Invest. New Drugs, 21(2):183–194, 2003.

[9] J E Latz, M O Karlsson, J J Rusthoven, A Ghosh, and R D Johnson.
A semimechanistic-physiologic population pharmacokinetic/pharmacodynamic model for neutropenia following pemetrexed therapy.
Cancer Chemother. Pharmacol., 57(4):412–426, 2006.

References III

[10] I F Troconiz, M J Garrido, C Segura, J M Cendros, P Principe, C Peraire, and R Obach.
Phase I dose-finding study and a pharmacokinetic/pharmacodynamic analysis of the neutropenic response of intravenous diflomotecan in patients with advanced malignant tumours.
Cancer Chemother. Pharmacol., 57(6):727–735, 2006.

[11] S J Kathman, D H Williams, J P Hodge, and M Dar.
A bayesian population PK-PD model of ispinesib-induced myelosuppression.
Clin. Pharmacol. Ther., 81(1):88–94, 2007.